Destined to Deceive: The Need to Regulate Deepfakes with a Foreseeable Harm Standard

Matthew D. Weiner

Follow this and additional works at: https://repository.law.umich.edu/mlr

Part of the Election Law Commons, Science and Technology Law Commons, and the Torts Commons

Recommended Citation
Available at: https://repository.law.umich.edu/mlr/vol122/iss4/5

https://doi.org/10.36644/mlr.122.4.destined

This Note is brought to you for free and open access by the Michigan Law Review at University of Michigan Law School Scholarship Repository. It has been accepted for inclusion in Michigan Law Review by an authorized editor of University of Michigan Law School Scholarship Repository. For more information, please contact mlaw.repository@umich.edu.
NOTE

DESTINED TO DECEIVE: THE NEED TO REGULATE
DEEPFAKES WITH A FORESEEABLE HARM STANDARD

Matthew D. Weiner*

Political campaigns have always attracted significant attention, and politicians have often been the subjects of controversial—even outlandish—discourse. In the last several years, however, the risk of deception has drastically increased due to the rise of “deepfakes.” Now, practically anyone can make audiovisual media that are both highly believable and highly damaging to a candidate. The threat deepfakes pose to our elections has prompted several states and Congress to seek legislative remedies that ensure recourse for victims and hold bad actors liable. These recent attempts at deepfake laws are open to attack from two different loci. First, there is a question as to whether these laws unconstitutionally infringe on deepfake creators’ First Amendment rights. Second, some worry that these laws do not adequately protect against the most harmful deepfakes. This Note proposes a new approach to regulating deepfakes. By delineating a “foreseeable harm” standard, with a totality-of-the-circumstances test rather than a patchwork system of discrete elements, this Note addresses both major concerns. Not only is a foreseeable harm standard effective, workable, and constitutionally sound; it is also grounded in existing tort law. Moreover, a recent Supreme Court decision pertaining to false statements and the First Amendment, United States v. Alvarez, lends support to such a standard. Adopting this standard will combat the looming threat of politically oriented deepfakes while preserving the constitutional right to free speech.

TABLE OF CONTENTS

INTRODUCTION ........................................................................................................772
I. THE PROBLEM WITH POLITICAL DEEPFAKES AND CURRENT
REGULATION ATTEMPTS ....................................................................................774
   A. Recent Examples of Politically Oriented Deepfakes .........................774
   B. Existing and Proposed Deepfake Laws .............................................779

*  J.D. Candidate, May 2024, University of Michigan Law School. I would like to thank my parents and siblings for their support throughout the writing process, Professor Don Herzog for pushing me to think critically about this problem and my proposed solution, and the many people that I discussed this idea with or that helped edit the piece. Any remaining mistakes are my own.
II. CONSTITUTIONAL AND POLICY PROBLEMS WITH CURRENT DEEPFAKE LAWS ................................................................. 783
   A. Constitutional Footing of Deepfake Laws ......................... 783
      1. Concerns from the Alvarez Court That Are Resolved by the Nature of Deepfakes ........................................ 785
         i. Counterspeech ...................................................... 785
         ii. “Other Legally Cognizable Harm” .......................... 786
         iii. Policing Truth and Chilling True Statements ............ 787
      2. Satisfying Scrutiny .................................................... 788
      3. Looming Issues ....................................................... 790
         i. Mens Rea ............................................................ 790
         ii. Context-Blind Restrictions ..................................... 791
         iii. Discriminatory Enforcement ................................. 792
   B. Shared Features of Existing Laws and Their Drawbacks ...... 793
      1. Disclaimers ............................................................ 793
      2. Subsequent Sharers .................................................. 795
      3. Reasonable Viewer .................................................. 796

III. THE BENEFITS OF A FORESEEABLE HARM APPROACH .......... 797
   A. Benefits of a Foreseeable Harm Standard ........................ 797
      1. Reasonable Viewer .................................................. 797
      2. Context-Blind Restrictions ....................................... 798
      3. Mens Rea ............................................................ 800
      4. Subsequent Sharers .................................................. 801
      5. Disclaimers ........................................................... 801
      6. Who and What Are Depicted? .................................... 802
   B. Support for a Foreseeable Harm Approach ........................ 803
      1. The Alvarez Concurrence .......................................... 804
      2. Other Applications of the “Reasonable Person” Standard ................................................................. 805

CONCLUSION .............................................................................. 807

INTRODUCTION

In the last several years, deepfakes—altered audiovisual media that can make anyone look like they are saying or doing things they have not said or
done—have infiltrated the internet.¹ Deepfakes are uniquely dangerous in two specific settings that have produced some of the most highly publicized and controversial examples. First, individuals can use “deep learning” technology to create fake pornographic content—a form of “revenge porn.”² Second, individuals can employ deepfakes in efforts to compromise political candidates and interfere with elections.³

As of October 1, 2023, nine states have adopted laws targeting one or the other of these two exigent threats.⁴ None of the enacted laws have faced challenges, but there is speculation that these laws may violate the First Amendment because they are overbroad to the point of chilling protected speech.⁵ Other states and Congress have proposed, but not yet adopted, similar laws. Though each jurisdiction’s approach to regulating deepfakes is different, there are common features that make these existing and proposed laws both constitutionally troubling and ineffective.

Scholars have been quick to offer policy and constitutional arguments in support of regulating revenge porn.⁶ However, laws targeting politically oriented deepfakes stand on more precarious ground, given the near-sacrosanct position that political discourse holds in our society.⁷ Thus, this Note turns its attention to political deepfakes: as Congress and state legislatures tread into

---


5. See VA. CODE ANN. § 18.2-386.2 (West 2023); TEX. ELEC. CODE ANN. § 255.004 (West 2023); CAL. ELEC. CODE § 20010 (West 2023); GA. CODE ANN. § 16-11-90 (West 2023); N.Y. PENAL LAW § 245.15 (McKinney 2023); FLA. STAT. § 775.0847 (2023); HAW. REV. STAT. § 711-1110.9 (2021); MINN. STAT. § 609.771 (2023); Act of July 23, 2023, ch. 360, 2023 Wash. Sess. Laws 1892 (to be codified at WASH. REV. CODE § 42).


these murky waters, it is critical to craft a constitutionally sound legislative response to political deepfakes.

This Note highlights the dangers of unrestricted political deepfakes and the shortcomings of current deepfake laws in an effort to chart a path forward for deterring them: a new framework for attaching liability to creators and publishers. Part I traces the recent proliferation of political deepfakes and the laws trying to keep pace. Part II demonstrates why deepfake laws serve compelling government interests, are necessary to curb harm, and do not unconstitutionally chill speech. It also examines the constitutional obstacles to and policy problems with existing deepfake laws. Part III proposes a simplified approach, arguing for a “foreseeable harm” standard for liability, considered in light of the totality of the circumstances. By avoiding over- and underinclusive tests for liability that can lead to unnatural distinctions among deepfakes, compelled speech, and invalid time, place, and manner restrictions, this standard will ensure that deepfake laws are narrowly tailored to regulate only the most harmful deepfakes and the most culpable perpetrators. Furthermore, deploying a foreseeable harm standard is not a far cry from current practice: it simply shifts the “reasonable person” standard from the deepfake viewer’s perspective to that of the creator or sharer, and Congress has already considered a similar approach, though not in full.

I. THE PROBLEM WITH POLITICAL DEEPFAKES AND CURRENT REGULATION ATTEMPTS

The threat of misinformation has taken center stage in recent elections, and perhaps the most effective form of misinformation is falsified audiovisual media. Technology that creates misleading audiovisuals continues to advance, posing an ever-growing threat to the integrity of our elections. This Part examines how deepfakes affect the political realm, the complications that stand in the way of solutions, and the attempted legislative solutions thus far.

A. Recent Examples of Politically Oriented Deepfakes

Over the last several years, the world has been exposed to a wide variety of deepfakes, with differing levels of sophistication and differing objectives. A

---


few key examples illuminate both the capabilities of deep learning technology and the stakes at play.

In 2019, the Massachusetts Institute of Technology’s “In Event of Moon Disaster” project released a video of President Richard Nixon announcing to the country that the Apollo 11 mission had failed, leaving the astronauts onboard stranded on the moon.\textsuperscript{12} This startlingly believable video, created as a learning tool on deepfakes, depicted an event that never occurred: MIT used “deep learning” technology coupled with AI software to simulate Nixon’s movements and match the synthetic video to the dialogue.\textsuperscript{13}

In 2020, a video of then-Speaker of the House Nancy Pelosi circulated on the internet. In the video, she appeared to be intoxicated.\textsuperscript{14} The deepfake did not alter the words Pelosi spoke, but it slowed down and distorted the audio itself to give the impression that she was inebriated.\textsuperscript{15} This time, the video was not a learning tool—instead, it was quickly circulated by political activists and viewed over 2.5 million times within the first few days of publication.\textsuperscript{16}

As tensions grew in 2022 in the ongoing armed conflict between Russia and Ukraine, Ukrainian President Volodymyr Zelenskyy became the target of a deepfake campaign.\textsuperscript{17} A hacked Ukrainian broadcasting service’s website briefly contained a video that seemed to depict Zelenskyy telling Ukrainian soldiers to surrender to Russian forces.\textsuperscript{18} In fact, Zelenskyy’s face was technologically inserted into the video, with his mouth movements manufactured to match the overlaid dialogue.\textsuperscript{19} Though the deepfake had flaws, experts still

\begin{thebibliography}{99}
\bibitem{13} Id.
\bibitem{15} Id.
\bibitem{18} Id.
\bibitem{19} See id.
\end{thebibliography}
worry about the impact it could have—and in fact already has had—on viewers. The video quickly reached several social media outlets, prompting a response from Zelenskyy himself, and it is possible that “lower-quality versions of the video could take on a life of their own in other parts of the world.”

The above examples are all instances of political deepfakes. Though definitions of deepfakes differ slightly, they can generally be defined as synthetically modified photographs or videos that appear to depict events that did not, in reality, occur. The possibilities of what a deepfake can depict are endless. Within the political realm, deepfake creators can insert a candidate somewhere they were not; edit real, preexisting video of a candidate to depict people or objects that were not actually present in the original; force a candidate to say words that they never said; or construct and place a candidate in an entirely fictitious scenario.

Of course, lies about politicians are not new: false depictions of politicians saying or doing damaging things date back to the nineteenth century. Until recently, these false depictions were more easily sniffed out. Even now, poorly made A.I.-generated media may still fail to deceive viewers. In recent years, though, deepfake technology has reached new levels of sophistication, making it increasingly difficult to detect which images are real and which are

---

20. These include both convincing viewers of this video’s veracity as well as raising doubts about any future video of Zelenskyy or other similar officials. Id.
21. The Ukrainian government had previously released a statement warning its citizens about this sort of deepfake. Id.
22. Id.
24. For further examples, see Shannon Bond, It Takes a Few Dollars and 8 Minutes to Create a Deepfake. And That’s Only the Start, NPR (Mar. 23, 2023, 5:00 AM), https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-start [perma.cc/UYSJ-QHEG].
26. Kamarck, supra note 25. For a discussion of the importance of preserving political cartoons and similar critical speech in relation to a defamation claim brought by a public figure, see Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 53–55 (1988). Notably, though these caricaturized depictions are meant to express specific sentiments about public figures, they do not masquerade as real depictions.
The examples above were highly publicized, making their falseness more readily discoverable; however, many deepfakes are not widely reported on, leaving individual viewers to decide for themselves what is real. Not only will some viewers believe these false videos, but their existence also undermines the credibility of real videos due to the lingering possibility that a nefarious user manipulated them. Thus, deepfakes’ increasing presence may lead to the delegitimization of news outlets, and viewers may selectively believe misinformation tailored to their preconceived notions, leading to further entrenchment of their views.

Hand in hand with these concerns is the threat political deepfakes pose to our elections. A deepfake creator can seek to “damage the reputation of [a candidate], incite a political base, or undermine trust in the election process.” Creators can also use deepfakes to shape policy discussions, implant fake evidence into the popular discourse on a variety of topics, and stir fierce rhetoric. Creators can even use deepfakes to directly extort politicians or fraudulently

28. ANAND & BIANCO, supra note 2.
29. Thomas Nygren, Mona Guath, Carl-Anton Werner Axelsson & Divina Frau-Meigs, Combatting Visual Fake News with a Professional Fact-Checking Tool in Education in France, Romania, Spain and Sweden, INFORMATION, May 2021, at 1. Even highly publicized deepfakes are effective, as supporters of the message conveyed will still hold the image to be indicative of the truth, even after it has been discredited. See Drew Harwell, Doctored Images Have Become a Fact of Life for Political Campaigns. When They’re Disproved, Believers ‘Just Don’t Care.’, WASH. POST (Jan. 14, 2020, 7:00 AM), https://www.washingtonpost.com/technology/2020/01/14/doctored-political-images[perma.cc/RBG5-E9VS] (describing how polarization allows people to insulate themselves from other narratives); see also Gerald G. Ashdown, Distorting Democracy: Campaign Lies in the 21st Century, 20 WM. & MARY BILL RTS. J. 1085, 1092–94 (2012) (describing the harms of unregulated, false campaign speech).
30. Jack Langa, Note, Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes, 101 B.U. L. REV. 761, 767 (2021). Langa, writing after the passage of Texas’s and California’s deepfake laws, similarly explored the danger of political deepfakes and how best to legislate against them. In addition to incorporating recent legislative developments, this Note also proposes a different solution. Though Langa highlights the importance of “likelihood to bring about [harm]” in assigning liability, id. at 787, he advocates for many of the features that this Note explicitly condemns, such as mandatory disclaimers, a reasonable viewer standard, and temporal cutoffs for liability. Id. at 789; see infra Part II.
32. Cote, supra note 31; Chesney & Citron, supra note 31, at 1768. This is further expounded infra Section II.B.3.
34. DEPT OF HOMELAND SEC., supra note 1.
35. See Mak & Temple-Raston, supra note 33.
gather information from confidential sources.\textsuperscript{36} Though these fears have not yet been fully realized, such persuasive misinformation poses a danger that will likely come to bear on future elections, especially because interfering with the democratic process may be the point.\textsuperscript{37}

For these reasons, mitigating the possible impacts of deepfakes on our elections should be a priority. Yet, techniques for detecting deepfakes have struggled to keep pace with the countervailing techniques to avoid detection, and any long-term technological solution remains elusive.\textsuperscript{38} Moreover, the time it takes to deploy systems to detect a deepfake often precludes them from being effective on a large scale.\textsuperscript{39} Therefore, we should expect deepfakes to remain a part of our lives going forward; the best we can hope for are new legislative solutions that deter bad actors, ensure recourse for victims, and protect election integrity.

Effective, workable legislation that assigns liability to deepfake creators and distributors is also essential because internet platforms are not in a position to regulate deepfakes. Not only does Section 230 of the Communications Decency Act dictate that platforms cannot be held liable for users’ posts,\textsuperscript{40} but there is also now a possibility that internet companies may not be permitted to police their own platforms of their own volition. In the recent case of \textit{NetChoice v. Paxton}, a group of large internet and technology businesses challenged a Texas statute that prohibited them from removing posts based on the viewpoint expressed or represented.\textsuperscript{41} The internet companies argued that they were entitled under the First Amendment to engage in editorial discretion and thus ought to retain control over “whether, to what extent, and in what manner to disseminate third-party-created content to the public.”\textsuperscript{42} The Fifth

\textsuperscript{36} KELLEY M. SAYLER, CONG. RSCH. SERV., IF11333, DEEP FAKEs AND NATIONAL SECURITY (2022).

\textsuperscript{37} Mak & Temple-Raston, supra note 33.

\textsuperscript{38} See Cristian Vaccari & Andrew Chadwick, \textit{Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News}, SOC. MEDIA & SOC’y, Jan.–Mar. 2020, at 1, 3; see also John A. Barrett, Jr., \textit{Free Speech Has Gotten Very Expensive: Rethinking Political Speech Regulation in a Post-Truth World}, 94 ST. JOHN’S L. REV. 615, 635–36 (2020). Even when deepfakes are proven inaccurate, trying to discredit a post cannot always “outpace” the original post, and reporting on a deepfake might actually fuel the image’s dissemination. Harwell, supra note 29.


\textsuperscript{40} 47 U.S.C. § 230(c)(1).

\textsuperscript{41} NetChoice, LLC v. Paxton, 49 F.4th 439, 444 (5th Cir. 2022). The platforms remain able to remove content in certain circumstances, such as if it directly incites criminality. \textit{Id.} at 452.

\textsuperscript{42} \textit{Id.} at 490 (quoting NetChoice v. Att’y Gen., 34 F.4th 1196, 1212 (11th Cir. 2022) (finding that a Florida law restricting platforms’ ability to moderate content violated the platforms’ First Amendment rights)).
Circuit disagreed and considered the companies’ actions to be censorship unprotected by the Constitution.\footnote{NetChoice, 49 F.4th at 494.} The Fifth Circuit’s holding raises questions about the future of content moderation and severely weakens the ability of internet companies to respond to hate speech.\footnote{Nithin Venkatraman, NetChoice, L.L.C. v. Paxton: 5th Circuit Sets Up Supreme Court Battle Over Content Moderation Authority of Social Media Giants, JOLT DIGEST (Oct. 21, 2022), https://jolt.law.harvard.edu/digest/netchoice-l-l-c-v-paxton-5th-circuit-sets-up-supreme-court-battle-over-content-moderation-authority-of-social-media-giants [perma.cc/HZB6-H6CP].} Though deepfakes did not play an explicit part in the Fifth Circuit’s opinion, the implications of this holding will make it difficult for social media platforms to combat deepfakes, even those that could have severe repercussions for our democratic system. If internet platforms cannot prevent dangerous misinformation from spreading, then it is critical that the legal system discourage this behavior and punish bad actors.

B. Existing and Proposed Deepfake Laws

Several states—California, Georgia, New York, Texas, Virginia, Florida, Hawaii, Minnesota, and Washington—have enacted bills targeting the dissemination of deepfakes.\footnote{CAL. ELEC. CODE § 20010 (West 2023); GA. CODE ANN. § 16-11-90 (West 2023); N.Y. PENAL CODE § 245.15 (McKinney 2023); TEX. ELEC. CODE ANN. § 255.004(d) (West 2023); VA. CODE. ANN. § 18.2-386.2 (West 2023); FLA. STAT. § 775.0847 (2023); HAW. REV. STAT. § 711-1110.9 (2021); MINN. STAT. § 609.771 (2023); Act of July 23, 2023, ch. 360, 2023 Wash. Sess. Laws 1892 (to be codified at WASH. REV. CODE § 42). Although Georgia, New York, Virginia, Florida, and Hawaii have passed laws targeting certain deepfakes, none of them currently have statutes specifically related to politically oriented deepfakes.} Legislators in other states and the U.S. House of Representatives have proposed similar statutes as well. This Section describes their key features.

In defining what constitutes a deepfake, one way in which these statutes differ is with respect to the type of media (audio, visual, or both) and the person depicted (either candidates only, or anyone). A bill proposed in Illinois identifies a type of media as a deepfake if a creator either digitally \textit{adds} a \textit{candidate} to an image they were not a part of, digitally \textit{alters} an image of a candidate to add another person, or “intentionally manipulate[s]" audio or visual media that portrays a candidate’s “appearance, speech, or conduct.”\footnote{S.B. 3746, 101st Gen. Assemb. Reg. Sess. (Ill. 2020). This bill, which would have amended ILL. ELEC. CODE § 6, never made it out of committee. Bill Status of SB3746: 101st General Assembly, ILLINOIS GEN. ASSEMBLY, https://www.ilga.gov/legislation/BillStatus.asp?GA=101&DocTypeID=SB&DocNum=3746&GAID=15&SessionID=108&LegID=125786 [perma.cc/4CJE-2LZB]. It suffered a similar fate after being reintroduced at the start of the next General Assembly in 2021. Bill Status of SB1717: 102nd General Assembly, ILLINOIS GEN. ASSEMBLY, https://www.ilga.gov/legislation/BillStatus_pf.asp?DocNum=1717&DocTypeID=SB&LegID=&GAID=16&SessionID=110&GA=102 [perma.cc/G636-35EM].} California’s and Minnesota’s statute and a bill proposed in New Jersey identify “deepfakes” as any audio or visual media that depict a candidate doing or saying...
something that they did not actually do or say.47 “Media” here means either technologically altered versions of real audiovisuals or technologically created audiovisuals “substantially derivative” of real audiovisuals.48 Texas’s statute defines “deepfakes” still differently, as videos that falsely depict a person (rather than only a candidate) doing something that they did not do.49 Washington’s statute characterizes a “deepfake” as an image, video, or audio depicting an individual’s appearance, speech, or conduct that has been intentionally manipulated with the use of generative adversarial network techniques or other digital technology in a manner to create a realistic but false [depiction].50 Finally, the proposed House bill’s definition of “media” encapsulates all audiovisual, visual, and audio records, created without consent, that depict a living or dead person engaged in a “material activity” that they did not engage in.51 Based on these definitions, the House bill would cover the greatest number of deepfakes while Illinois’s would likely reach the fewest.

The liability standard also varies from statute to statute, based on criteria such as how convincing the deepfake is, the publisher’s intent, and the probability of harm. Under Washington’s, California’s, and Minnesota’s respective statutes, and the proposed statutes in New Jersey and Illinois, liability attaches if a reasonable viewer would believe the audiovisual to be authentic and if their “understanding or impression” of its “expressive content” would be different than if they had seen an unaltered audiovisual.52 Liability, then, depends on a reasonable viewer rather than a reasonable creator. The proposed House bill would use a reasonable viewer standard, with the added requirement that the deepfake must also be “substantially likely” to improperly interfere with an election when the deepfake depicts a deceased person.53 By contrast, in Texas, a video created with the “intent to injure a candidate or influence the result of an election” may lead to liability for its creator, the susceptibility of the viewer notwithstanding.54


49. ELEC. § 255.004(e).


51. DEEPFAKES Accountability Act, H.R. 5586, 118th Cong. § 1041(a)(2)(A)–(C), (g)(1) (2023). This bill was introduced by Representative Yvette Clarke; previous iterations were introduced by Representative Clarke in 2019 and 2021. DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019); DEEP FAKES Accountability Act, H.R. 2395, 117th Cong. (2021). It is beyond the scope of this Note to address this element in its entirety, but the House’s definition is the best suited to adapt to dangerous deepfakes.


54. ELEC. § 255.004(d) (emphasis added).
The statutes also differ with respect to what actions (creating or distributing) can result in liability and whether intent is an element. They also vary as to when liability can attach for those actions (in the run-up to an election or at any time). Texas’s statute limits liability to a person who creates a deepfake and “causes the [deepfake] to be published” in the run-up to an election.\(^{55}\) New Jersey’s proposed statute and California’s and Minnesota’s statutes limit liability to those who distribute a deepfake in the run-up\(^{56}\) to an election with the intent of influencing that election or harming a candidate’s reputation.\(^{57}\) Illinois’s proposed statute would assign liability to those who produce or publish a deepfake, with no timeframe specified.\(^{58}\) Under the proposed House bill, criminal liability attaches to any person who produces a deepfake with intent to “interfere in an official proceeding, including an election, provided the advanced technological false personation record did in fact pose a credible threat of instigating or advancing such”; civil liability is limited to any person who knowingly removes or “meaningfully obscure[s]” a disclaimer to a deepfake for the purposes of influencing an election, or anyone who produces a deepfake without the requisite intent for criminal liability—with no timeframe specified.\(^{59}\) The presence or absence of a temporal cutoff, as well as whether subsequent sharers of a deepfake can be held liable, greatly affects the number of deepfakes and perpetrators that a law can curtail.

Several of these statutes also incorporate disclaimer requirements to address concerns about the believability and resulting impact of deepfakes. Even these requirements, though, differ in scope. The current and proposed statutes in four states—Washington, California, Illinois, and New Jersey—require that any visual deepfake include an unambiguous textual disclaimer for the duration of the deepfake; any audial deepfake must include an unambiguous audial disclaimer at the beginning, end, and—if the audio is a certain length (two minutes or longer)—throughout the deepfake.\(^{60}\) The 2021 version of the House bill required that any visual deepfake continuously display an unambiguous textual disclaimer; that any audial deepfake include an unambiguous audial disclaimer at the beginning, end, and—if the audio is a certain length (two minutes or longer)—throughout the deepfake.\(^{60}\) The 2021 version of the House bill required that any visual deepfake continuously display an unambiguous textual disclaimer; that any audial deepfake include an unambiguous audial disclaimer, as well as an additional disclaimer if the deepfake exceeds two minutes; that any deepfake containing both audio and video components

\(^{55}\) Id. at (d)(1)–(2).

\(^{56}\) New Jersey’s proposed statute and California’s statute use sixty days as a temporal cutoff for assigning liability. N.J. Assemb. B. 4985; CAL. ELEC. CODE § 20010(a) (West 2023). Texas’s law, on the other hand, is limited to thirty days before an election. ELEC. § 255.004(d)(2). Finally, Minnesota’s statute applies to deepfakes disseminated ninety days before an election. MINN. STAT. § 609.771(2)(1) (2023).

\(^{57}\) N.J. Assemb. B. 4985; CAL. ELEC. CODE § 20010(a) (West 2023); MINN. § 609.771(2)(3).


\(^{59}\) See DEEPFAKES Accountability Act, H.R. 5586, 118th Cong. § 1041(f)(1)–(2) (2023) (emphasis added). It is beyond the scope of this Note to consider this element in its entirety, but the House bill would most effectively and constitutionally address different actors’ conduct.

display both an audial and unambiguous textual disclaimer; and that any deepfake containing a “moving visual element” display an “embedded digital watermark” disclaimer, distinct from the textual disclaimer.\textsuperscript{61} Texas’s statute does not contain disclaimer requirements.\textsuperscript{62} Not only do more intricate disclaimers, such as digital watermarks, reduce the likelihood that viewers will be deceived by a deepfake, but they also make it more difficult for ill-intentioned sharers to remove them.\textsuperscript{63}

Statutes also diverge as to whether they prescribe civil or criminal liability. California’s and Washington’s statutes, as well as New Jersey’s and Illinois’s proposed statutes, assign civil liability through private causes of action.\textsuperscript{64} Candidates who are the subject of a deepfake may seek damages under all four statutes. California’s statute and Illinois’s proposed statute allow for “any registered voter” to seek injunctive relief.\textsuperscript{65} Texas’s and Minnesota’s statutes assign criminal liability.\textsuperscript{66} As noted above, the proposed House bill contains both criminal and civil liability provisions, categorizing the production of and intent to distribute a deepfake for the purposes of influencing an election as a criminal offense, whereas deepfake producers with lesser mens rea and culpable subsequent sharers are subject to only civil liability.\textsuperscript{67} The private right of action available under the proposed House bill is limited to a person or entity that a deepfake actually depicts, with no recourse available to registered voters harmed by viewing the deepfake.\textsuperscript{68} In addition to affecting the type of accountability that a defendant can face, this difference also impacts who can seek recourse for their injuries and how they can do so.

\textsuperscript{61} H.R. 2395 § 1041(b)–(e). The bill states that “[n]ot later than 1 year after the date of enactment of this section, the Attorney General shall issue rules governing the technical specifications of the digital watermarks required.” Id. § 1041(k)(3). The 2023 version of the bill, however, removes specific language about watermarks. See H.R. 5586 § 1041.

\textsuperscript{62} TEX. ELEC. CODE ANN. § 255.004 (West 2023). As developed \textit{infra} Section II.B.1, any strict disclaimer requirement carries with it both constitutional and policy-based concerns.

\textsuperscript{63} Langa, \textit{supra} note 30, at 789.

\textsuperscript{64} CAL. ELEC. § 20010(c)(1)–(3); N.J. Assemb. B. 4985; 2023 Wash. Sess. Laws 1892; Ill. S.B. 3746.

\textsuperscript{65} N.J. Assemb. B. 4985; 2023 Wash. Sess. Laws 1892; CAL. ELEC. § 20010(c)(1)–(2); Ill. S.B. 3746. California’s current statute, effective until January 1, 2027, does not contain a private right of action for voters. ELEC. § 20010(c). Beginning January 1, 2027, the section of the code creating this private right of action will take effect. Id. § 20010(c)(1), (e). As any suit must still satisfy constitutional standing requirements, plaintiffs will still have to demonstrate an “injury in fact” that bears a causal connection to the deepfake and may be redressed by the requested relief. See \textit{Lujan v. Defs. of Wildlife}, 504 U.S. 555, 560–61 (1992).

\textsuperscript{66} TEX. ELEC. CODE ANN. § 255.004(c) (West 2023); MINN. STAT. § 609.771(2) (2023).

\textsuperscript{67} DEEPFAKES Accountability Act, H.R. 5586, 118th Cong. § 1041(f)–(g) (2023). Minnesota’s statute makes injunctive relief, but not civil damages, available against “any person who is reasonably believed to be about to violate or who is in the course of violating” the law; government attorneys, individuals portrayed by a deepfake, and candidates that are or may be hurt by a deepfake may seek injunctive relief. \textit{Minn. Stat.} § 609.771(4) (2023). There are benefits and drawbacks to both civil and criminal liability schemes, as described \textit{infra} Section II.A.I.iii; it is beyond the scope of this Note to make a recommendation for future legislation.

\textsuperscript{68} Id. § 1041(g).
As detailed above, there are many similarities across existing legislation targeting politically oriented deepfakes. Common among them is a patchwork system of discrete, dualistic elements, including: whether the deepfake includes the requisite form of disclaimer, whether the defendant was the original publisher or a subsequent sharer, whether the deepfake portrays a candidate or another person, whether the deepfake was posted in the run-up to an election, and whether the defendant acted with the requisite mens rea. Each statute picks a different combination from this menu of on-off switch provisions. Part II highlights the problems with each of these tests—especially in how they interact with each other.

II. CONSTITUTIONAL AND POLICY PROBLEMS WITH CURRENT DEEPFAKE LAWS

This Part demonstrates the need for a new approach to regulating politically oriented deepfakes. It examines common First Amendment concerns with policing deepfakes and distinguishes the speech at issue from that in United States v. Alvarez, a recent and influential Supreme Court case about an attempt to prohibit “stolen valor.” As this Part argues, Xavier Alvarez’s lie was less harmful and more easily disproved than the shadowy, far-reaching lies spread by deepfakes. This Part then assesses how the Alvarez Court’s concerns map onto deepfake laws’ shared features, pointing out shortcomings of existing laws and proposals.

A. Constitutional Footing of Deepfake Laws

It is no secret that any limitation on free speech will face opposition.\(^69\) Hesitance to “chill” protected speech is particularly heightened for political speech,\(^70\) even decidedly false speech.\(^71\) In United States v. Alvarez, the Supreme Court found unconstitutional a statute that criminalized falsely claiming to have received the Congressional Medal of Honor.\(^72\) Xavier Alvarez was indicted under the Stolen Valor Act for holding himself out as a Medal of Honor recipient at a public meeting.\(^73\) Congress had not in fact awarded Alvarez the

---


70. See Snyder v. Phelps, 562 U.S. 443, 452–53 (2011) (noting that “speech on public issues . . . is entitled to special protection” and describing public issues as “any matter of political, social, or other concern to the community” (internal citations omitted) (quoting Connick v. Myers, 461 U.S. 138 (1983))).

71. See Rickert v. State, Pub. Disclosure Comm’n, 168 P.3d 826 (Wash. 2007) (finding that a political candidate could not be held liable even for false statements about her competitor).


73. Id. at 713–14.
Medal of Honor, and the Court acknowledged that his statement was nothing but an “intended, undoubted lie.”

Alvarez challenged his indictment on the grounds that the Stolen Valor Act violated his First Amendment right to free speech, and the Court agreed. The purpose of the Act was to safeguard the reputation of the Medal of Honor, an award created “so the Nation could hold it in its highest respect and esteem” individuals who defended the safety of this country “with extraordinary honor.” The Court stated this was “a legitimate Government objective” that Congress was “right and proper” to pursue. Nevertheless, the Court concluded that, when subjected to the “sometimes inconvenient principles of the First Amendment,” the Act impermissibly infringed upon Alvarez’s constitutional rights. “[C]ontent-based restrictions on speech [must] be presumed invalid,” and due to the “probable, and adverse, effect of the Act on freedom of expression,” the Court held that the government had not overcome this presumption. Because there was no principle limiting when the government could punish defendants for making such false statements, the government’s “censorial power” would chill the fundamental freedoms of “speech, thought, and discourse,” which are necessary to the operation of a democracy.

Although the Alvarez Court did not say that all false statements fall within the ambit of the right to free speech, the Court did consider several constitutional obstacles that must be overcome before limiting even false speech. Some of the obstacles that the justices raised arguably do not apply to deepfakes, given their unique nature and characteristics. Still, some of the constitutional limits loom large in the debate surrounding current deepfake laws.

---

74. Id. at 714–15.
75. Id. at 714.
76. Id. at 715.
77. Id.
78. Id. at 715–16.
79. Id. at 716–17, 722–23 (quoting Ashcroft v. ACLU, 542 U.S. 656, 660 (2004)).
80. Id. at 723.
81. Rather, the Court recognized that “there are instances in which the falsity of speech bears upon whether it is protected” and merely rejected “the notion that false speech should be in a general category that is presumptively unprotected.” Id. at 721–22.
82. Id. at 725–29 (analyzing when speech restrictions may be constitutional and how the Stolen Valor Act mapped onto these considerations).
1. Concerns from the Alvarez Court That Are Resolved by the Nature of Deepfakes

   i. Counterspeech

   Justice Kennedy, writing for a plurality of the Court, offered what has become a common argument for disallowing limitations on speech: “The remedy for speech that is false is speech that is true.” The “counterspeech doctrine” traces its roots back to Justice Brandeis’s concurring opinion in *Whitney v. California*. Whitney, a member of the Communist Labor Party, was convicted of assembling to advocate for a violent overthrow of the government. Despite concurring with the outcome—which *Brandenburg v. Ohio* later overruled—Justice Brandeis argued that free speech extended to critiques of the government, even those that may be proven untrue. Justice Brandeis wrote that “[i]f there be time to expose through discussion the falsehoods and fallacies . . . the remedy to be applied is more speech, not enforced silence.”

   In *Alvarez*, the Court built on this reasoning and stated that, if Congress’s interest in passing the Stolen Valor Act was to make clear who had been awarded military honors, the government could maintain a database that lists every recipient. This, the Court said, would counteract fictional claims like Alvarez’s. However, while the counterspeech doctrine constitutionally protected Alvarez’s false statements, such counterspeech is beyond the realm of possibility for deepfakes. While it would be hardly conceivable for the government to maintain a database of every portrayal of a candidate as a fact-checking backstop, it is unfathomable for it to maintain a database of every image or video into which a candidate could possibly be inserted. Therefore, there is simply no way to ensure the accuracy of an image or video in the same way that the Court envisioned in *Alvarez*.

---

83. *Id.* at 727.
88. *Id.* at 377.
90. *Id.*
92. These differences suggest that there is no less restrictive means for combatting deepfakes; therefore, civil or criminal penalties are necessary. Protecting the integrity of our elections also satisfies the “compelling interest” element of strict scrutiny. *See infra* Section II.A.2. It is
In other ways, too, counterspeech is not nearly as effective in defending against deepfakes. First, the speed with which deepfakes can spread makes it difficult to adequately rebut each misrepresentation. Second, people are more likely to believe a deepfake depicts the truth than they are other false statements. As the saying goes: seeing is believing. Moreover, even when confronted with a deepfake’s inconsistencies, our brains resist entirely discounting its reality. There are thus serious obstacles to effective counterspeech in the deepfake context.

Additionally, the Alvarez Court raised the role of public backlash in countering false speech: Alvarez was ridiculed for lying, and other false claimants would be too. According to the Court, this public condemnation would serve to reduce the deceptive impact of misinformation. Alvarez spread false claims about himself; his speech was visibly connected with him. Deepfake creators and subsequent sharers, however, often keep their identities secret, with the help of sophisticated technology. The creator of a deepfake—disguised by a false name or no name at all, hidden behind layers of sophisticated protection from identification, and empowered by subsequent sharers that need not attach the creator’s name to their own posts—will often not face disrepute in the way the Court imagined.

ii. “Other Legally Cognizable Harm”

In Alvarez, the government argued that there was no First Amendment protection for false statements. The Court disagreed: the state could not condemn dishonesty alone. Liability for false statements requires the presence

---

96. Sunstein, supra note 91, at 422. This may be attributable to people’s tendency to be overly trusting of a video’s accuracy, even in the face of clear warning signs to the contrary. Köbis et al., supra note 95, at 10–11 (finding that people’s bias toward authenticity led them to guess that nearly 70 percent of videos were authentic even after being told that only 50 percent were authentic).
98. Id. at 713, 727.
99. Delfino, supra note 7, at 899.
100. Alvarez, 567 U.S. at 719.
of “defamation, fraud, or some other legally cognizable harm associated with a false statement.”\textsuperscript{101} Political deepfakes cause just that. They cannot be considered “falsity and nothing more.”\textsuperscript{102} At best, they are falsities accompanied by reputational damage to a candidate. At worst, they are falsities that unduly influence an election. Either way, deepfakes fall squarely within the examples that the Court distinguished from Alvarez’s protected speech.\textsuperscript{103}

### iii. Policing Truth and Chilling True Statements

Beyond Alvarez, First Amendment principles limiting government policing of speech are inapplicable to deepfakes. An oft-cited concern with any limitation on First Amendment protections is that the government should not be the “arbiter of truth.”\textsuperscript{104} This consideration was a major factor in Rickert v. State, Public Disclosure Commission, in which the Washington Supreme Court did not extend liability to a political candidate who made false statements about her opponent.\textsuperscript{105} The Court was hesitant to “assume[] that the government is capable of correctly and consistently negotiating the thin line between fact and opinion in political speech.”\textsuperscript{106} If the government could censor statements it determined to be false, there could be partisan manipulation of “facts.”\textsuperscript{107} Greater restrictions on deepfakes, however, do not implicate this concern. By definition, deepfakes present inaccurate information and, therefore, are not even arguably factual.\textsuperscript{108} For the same reason, there is no concern that statutes aimed at curbing deepfakes will chill individuals from making or

---

\textsuperscript{101} See id.
\textsuperscript{102} Id.
\textsuperscript{103} See Jessica Ice, Note, Defamatory Political Deepfakes and the First Amendment, 70 CASE W. RESR. L. REV. 417, 419, 437 (2019) (highlighting both the reputational damage to a candidate and their campaign and the “societal damages caused by the video if it is allowed to persist in the public sphere”). Others have argued that deepfakes fall into the existing misappropriation of likeness or invasion of privacy torts, see Zahra Takhshid, Retrievable Images on Social Media Platforms: A Call for a New Privacy Tort, 68 BUFF. L. REV. 139, 157 (2020), or that deepfakes are not speech at all, see generally Marc Jonathan Blitz, Deepfakes and Other Non-Testimonial Falsities: When Is Belief Manipulation (Not) First Amendment Speech?, 23 YALE J. & TECH. 160 (2020). It is beyond the scope of this Note to analyze these arguments in their entirety, but their success is unlikely. Kavyasri Nagumotu, Deepfakes Are Taking over Social Media: Can the Law Keep Up?, 62 IDEA: L. REV. FRANKLIN PIERCE CTR. FOR INTELL. PROP. 102, 128 (2022).
\textsuperscript{104} Alvarez, 567 U.S. at 752 (Alito, J., dissenting).
\textsuperscript{106} Id. at 829.
\textsuperscript{107} For other noted concerns, see Katrina Geddes, Ocularcentrism and Deepfakes: Should Seeing Be Believing?, 31 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1042, 1076 (2021). See also Sunstein, supra note 91, at 398 (noting that the government could also make genuine mistakes about what is true and what is false).
\textsuperscript{108} Ice, supra note 103, at 439.
publishing true statements.\textsuperscript{109} Chilled speech was at issue in New York Times v. Sullivan, the landmark First Amendment case that established “actual malice” as a requisite element of defamation claims brought by public figures.\textsuperscript{110} In that case, the Court made clear that First Amendment protections extend to “erroneous statements honestly made” as a result of “negligence or carelessness.”\textsuperscript{111} The Court emphasized the need for “breathing space” for free speech in general and for political speech specifically.\textsuperscript{112} Deepfakes, though, are decidedly not “honestly made.” Assigning liability to creators for harmful deepfakes does not carry the same danger of discouraging innocent speech.

2. Satisfying Scrutiny

The plurality advocated for an “exacting scrutiny” approach in Alvarez, requiring a compelling government interest to restrict speech.\textsuperscript{113} The restriction must also be narrowly tailored to serve that interest.\textsuperscript{114} Though the government’s interest in protecting the “value and meaning” of the Medal of Honor was compelling,\textsuperscript{115} the Court deemed the Stolen Valor Act unnecessary to serve this interest because there was no evidence that “the public’s general perception of military awards is diluted by false claims” of attainment.\textsuperscript{116} The other impacts of Alvarez’s and similar false claims—such as offending actual recipients of the Medal of Honor—also could not justify the speech restriction.

The same cannot be said for the impacts of deepfakes. First, deepfakes have already changed the public perception of candidates.\textsuperscript{117} As compared to the need to protect the “general perception of military awards,” the need to remedy defamatory impacts on a candidate is much more tangible. Second, deepfakes threaten to irreparably alter future elections. Notably, Justice Breyer’s concurrence in Alvarez, which plays a central role in supporting the standard proposed by this Note, recognized that false speech related to political campaigns is “more likely to make a behavioral difference” among viewers than are other forms of false speech.\textsuperscript{118} Thus, the necessity of legislation to

\begin{itemize}
\item \textsuperscript{109} It could be argued that the fear of inadvertently sharing a deepfake would chill subsequent sharers from reposting online content; however, as discussed infra, a lack of proof of reckless or intentional conduct in sharing a deepfake—such as altering or removing a disclaimer—would protect against inappropriately attaching liability. See infra Section III.A.4.
\item \textsuperscript{111} Sullivan, 376 U.S. at 278, 283 n.24.
\item \textsuperscript{112} Id. at 271–72.
\item \textsuperscript{113} See United States v. Alvarez, 567 U.S. 709, 715 (2012). Justice Breyer argued instead that intermediate scrutiny should apply. Id. at 725, 730 (Breyer, J., concurring).
\item \textsuperscript{114} Id. at 737–38 (Breyer, J., concurring).
\item \textsuperscript{115} Id. at 726 (majority opinion).
\item \textsuperscript{116} Id.
\item \textsuperscript{117} See Langa, supra note 30, at 766 n.28, 767.
\item \textsuperscript{118} Alvarez, 567 U.S. at 738 (Breyer, J., concurring).
\end{itemize}
protect this interest is much clearer for deepfakes than for the type of false speech in Alvare.

As to the interest itself, one could argue that there exists a freestanding compelling interest in ensuring free and fair elections.\textsuperscript{119} Thus, given the unprecedented threat that deepfakes pose to our elections, the government would be justified in taking steps to curb their deleterious impact. Yet, the Court has often snubbed this interest when it comes to limiting First Amendment rights. For example, in \textit{Citizens United v. FEC}, the Court did not recognize this broad interest in protecting elections. Rather, the Court made clear that “laws that burden political speech” will rarely and only narrowly pass constitutional muster.\textsuperscript{120} The government could not censor on the basis of speaker or viewpoint and could only restrict speech if it would interfere with the proper functioning of government entities.\textsuperscript{121} The documentary in question, which critiqued a presidential candidate, could not be construed to have such an effect and, therefore, deserved First Amendment protection.\textsuperscript{122}

Similar arguments will no doubt be made regarding deepfakes, given their ostensible lack of concrete interference with a specific government entity’s functioning. However, the potential for deepfakes to be used to extort politicians, gain access to confidential information, and infuse misinformation into policy debates presents real threats to any number of government functions, as employees may be prevented from doing their job or influenced to take negative action.\textsuperscript{123} Therefore, even if the Court does not think that ensuring free and fair elections is enough, alone, to justify First Amendment restrictions, political deepfakes could surpass even the high bar set by \textit{Citizens United}. It is also worth noting the inconsistency in the Court’s approach to “protecting” our elections: in order to prevent what it views as too much government influence on the electoral process, the Court is willing to allow malignant influences to fester when private actors are behind them.

The Court’s other reasons for protecting the speech in \textit{Citizens United} may actually support the case for regulating deepfakes. First, the Court cited \textit{Buckley v. Valeo} for the notion that “the ability of the citizenry to make informed choices” is essential to the operation of our democratic republic.\textsuperscript{124}

\begin{flushright}
\textsuperscript{119} \textit{Ex parte Stafford}, 667 S.W.3d 517, 525 (Tex. App. 2023) (recognizing a compelling interest in limiting the impact of fraudulent statements on elections and, therefore, the population in a case challenging a portion of Texas’s statute criminalizing certain campaign communications—separate from its deepfake provisions); see also Ice, \textit{supra} note 103, at 439; Langa, \textit{supra} note 30, at 781. For discussion of this compelling interest in the context of gerrymandering, see Daunt v. Benson, 956 F.3d 396, 426 (6th Cir. 2020) (Readler, J., concurring) (quoting Amicus Curiae Brief of Brennan Ctr. for Just. in Support of Defendant-Appellee and Affirmance at 24, \textit{Daunt}, 956 F.3d 396 (No. 19-2377)).
\textsuperscript{120} \textit{Citizens United v. FEC}, 558 U.S. 310, 340 (2010).
\textsuperscript{121} \textit{Id.} at 341.
\textsuperscript{122} \textit{Id.}
\textsuperscript{123} \textit{See supra} notes 33–34 and accompanying text.
\textsuperscript{124} \textit{Citizens United}, 558 U.S. at 339 (quoting Buckley v. Valeo, 424 U.S. 1, 14 (1976)).
\end{flushright}
Corporations, the Court said, may be uniquely well-suited to publicize important information and aid voters’ decisionmaking.\textsuperscript{125} Conversely, deepfakes expressly undermine informed decisionmaking. Deepfakes, by design, muddy the waters of public discourse. Second, the Court stated that “voters must be free to obtain information from diverse sources in order to determine how to cast their votes.”\textsuperscript{126} If this right to information is unconditional, it would appear to cut against any form of political speech limitation. But this right is not unconditional. For example, all fifty states have laws restricting electioneering activities near polling places, and the Court upheld Tennessee’s version of this law because the state had a compelling interest in “protecting voters from confusion and undue influence.”\textsuperscript{127} Evidently, the Court acknowledges that certain information sources harm, rather than help, voters. The danger of deepfakes is more similar to the danger posed by the electioneering activities in \textit{Burson} than the campaign funding in \textit{Citizens United}. In fact, deepfakes are an even clearer example of “confusion and undue influence”\textsuperscript{128} than campaigning at polling places, which is limited across the nation.

3. Looming Issues

i. Mens Rea

One constitutional obstacle that may still plague deepfake laws is the mens rea requirement. The Supreme Court has repeatedly pointed out that the cases in which it condemned false speech involved false statements made knowingly or with reckless disregard for the statement’s veracity.\textsuperscript{129} As discussed, the creator of a deepfake undoubtedly makes a false statement “knowingly.”\textsuperscript{130} But the mens rea of a subsequent sharer is less clear. If a statute assigns liability to everyone that shares a deepfake, it may enslave the “careless” speaker about whom the Court was worried.\textsuperscript{131} Rather, it is necessary to consider the individualized circumstances of each post in order to determine whether the sharer was aware of the consequences that would follow, and thus the sharer’s culpability.\textsuperscript{132}

\textsuperscript{125} Id. at 364.
\textsuperscript{126} Id. at 341.
\textsuperscript{128} Id. at 199.
\textsuperscript{129} Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52 (1988) (pertaining to fictional admissions that minister Jerry Falwell engaged in incestuous behavior with his mother in a nationally circulated magazine); United States v. Alvarez, 567 U.S. 709, 719 (2012).
\textsuperscript{130} Ice, \textit{supra} note 103, at 434.
\textsuperscript{132} \textit{See infra} Section III.A.3.
ii. **Context-Blind Restrictions**

In *Alvarez*, the Court was concerned that the Stolen Valor Act applied regardless of context: the Act criminalized “false statement[s] made at any time, in any place, to any person,” reaching false speech “in almost limitless times and settings.”\(^{133}\) Notably, the *Alvarez* Court’s concerns are distinguishable from the concerns animating the Court’s doctrine for traditional “time, place, and manner restrictions” (TPM) for speech in public fora.\(^{134}\) In *Ward v. Rock Against Racism*, Justice Kennedy—again writing for the Court—discussed requirements for the latter: the government may limit speech in a specific setting if the limitation (1) is content-neutral, (2) is narrowly tailored to serve a significant government purpose, and (3) leaves open alternative ways to communicate the speaker’s message.\(^{135}\) If the government wishes to prohibit speech in a public forum in a specific context, it must sufficiently justify the prohibition and cannot selectively enforce it.

Thus, in *Alvarez*, the Court expressed a different concern: that the statute’s “sweeping” breadth and *indifference to context* were in tension with First Amendment rights.\(^{136}\) Specifically, the Court noted that the Stolen Valor Act would treat statements made in a public meeting and “whispered conversations within a home” equally, suppressing both.\(^{137}\) Restrictions cannot be absolute and context-blind. They must be limited to the situations in which the speech would result in a concrete harm.\(^{138}\)

Some of the current drafts of deepfake legislation have attempted to comply with this constitutional prerequisite by imposing liability only during certain time periods, such as sixty days before a general election.\(^{139}\) This strategy treads into the realm of TPM restrictions discussed above.\(^{140}\) It also resembles previously invalidated provisions of statutes limiting political speech, like the Bipartisan Campaign Reform Act (BCRA), which prohibited corporate “electioneering communications” in the run-up to elections.\(^{141}\) Rigid temporal cut-offs like this draw arbitrary distinctions between punishable and

---


134. Though seemingly easy to distinguish, the differences between the two threads of concern may be relevant if large social media platforms are considered public spaces, as in *NetChoice, LLC v. Paxton*, 49 F.4th 429 (5th Cir. 2022). See * supra* Section I.A.


137. *Id.*

138. *Id.* at 723.

139. See Wilkerson, * supra* note 8, at 424.

140. See * supra* notes 134–135 and accompanying text.

141. Bipartisan Campaign Reform Act of 2002, Pub. L. No. 107-155, §§ 201–04, 116 Stat. 81, 88–92; *Citizens United v. FEC*, 558 U.S. 310 (2010). It has been argued that the fundamental differences between corporate-funded political speech and deepfakes render this similarity irrelevant. See Anna Pesetski, Note, *Deepfakes: A New Content Category for a Digital Age*, 29 WM. & MARY BILL RTS. J. 503, 518 (2020). However, this does not go far to suggest that the temporal
nonpunishable deepfakes: how does a deepfake posted sixty-one days before an election cause significantly less harm than a deepfake posted one day later, irrespective of their other features? A better strategy for regulating deepfakes is to avoid wading into the treacherous waters of TPM altogether. A focus on foreseeable harm rather than TPM can sidestep this constitutional bind.142

iii. Discriminatory Enforcement

Finally, an ongoing fear that accompanies any attempt to limit political speech is discriminatory enforcement.143 For deepfake regulations that assign criminal liability, one could argue that, because the determination of whether harm is foreseeable enough to trigger liability is unavoidably subjective, prosecutors will be able to choose whom to pursue charges against on a strictly partisan basis. This could, in turn, give rise to discriminatory enforcement of deepfake laws. However, greater oversight and sanctions can confine such prosecutorial misconduct.144 Additionally, assigning civil liability could mitigate this problem by creating a private right of action, which would allow victims of deepfakes—people who suffer either reputational or electoral harm caused by a deepfake—to seek justice for themselves.145 Undoubtedly, the parties to these lawsuits will fall along partisan lines (for example, a Democratic plaintiff suing a Republican deepfake distributor or vice versa), but the ability for all injured parties to bring a claim may help eliminate the risk of selective enforcement by the party in power.146 A private right of action does raise standing questions. Although candidates whose reputations are damaged by a deepfake have a straightforward injury in fact, voters who are harmed because of a deepfake’s larger impact on an election may face difficulty establishing an individualized injury as opposed to a generalized grievance.147 However, that is

limitations serve to remedy the Court’s fears unless deepfake laws are accepted more broadly as constitutional.

142. See infra Section III.A.2.

143. See e.g., Alvarez, 567 U.S. at 734 (Breyer, J., concurring).


146. Of course, these statutes may bring with them other complications, such as frivolous lawsuits against satirists and the like in order to curb messages that the claimants disagree with. Cf. Langa, supra note 30, at 798. If this standard is deployed correctly, however, claims should fail when they are not genuine.


There may also be concerns of partisan bias among judges and juries hearing these cases; perhaps the political persuasion of the defendant will impact the arbiter’s determination of liability or guilt.\footnote{149}{See, e.g., Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 55 (1988).} Of course, this is not the only category of cases that could implicate partisan bias.\footnote{150}{James P. Brady, Fair and Impartial Railroad: The Jury, the Media, and Political Trials, 11 J. Crim. Just. 241 (1983).} To a certain extent, bias is inevitable.\footnote{151}{See Shamena Anwar, Patrick Bayer & Randi Hjalmarsson, Politics in the Courtroom: Political Ideology and Jury Decision Making, 17 J. Eur. Econ. Ass’n 834 (2018) (describing the effect of partisan bias in Swedish juries, in which some of the jurors are politically affiliated officials).} Some have argued that the current safeguards of jury selection and appellate review are, if not satisfactory, the closest we can come to a solution.\footnote{152}{See Brady, supra note 150; Cassandra Burke Robertson, Judicial Impartiality in a Partisan Era, 70 Fla. L. Rev. 739 (2018).}

Additional procedural measures to protect against partisan bias include jury instructions on potential bias\footnote{153}{See Anona Su, A Proposal to Properly Address Implicit Bias in the Jury, 31 Hastings Women’s L.J. 79, 98–99 (2020).} and the exclusion of substantially prejudicial evidence.\footnote{154}{See generally David Sonenshein & Charles Fitzpatrick, The Problem of Partisan Experts and the Potential for Reform Through Concurrent Evidence, 32 Rev. Litig. 1 (2013) (proposing court-appointed, neutral experts akin to civil law jurisdictions or cooperation between competing experts).} Though it is likely implausible to entirely eliminate bias in politically charged trials, these existing safeguards allow legislatures to address harmful deepfakes and produce more neutral verdicts.

B. Shared Features of Existing Laws and Their Drawbacks

1. Disclaimers

One common point of discussion for deepfake liability is the effect of disclaimers. The presence of a disclaimer may drastically reduce the chance that a reasonable viewer is deceived.\footnote{155}{Ice, supra note 103, at 434.} Therefore, under a reasonable viewer standard, including a disclaimer that a video has been modified frees its creator of liability.\footnote{156}{Id.; Douglas Harris, Note, Deepfakes: False Pornography Is Here and the Law Cannot Protect You, 17 Duke L. & Tech. Rev. 99, 117 (2019).} Several states even explicitly waive liability for publishers that attach a disclaimer to a deepfake.\footnote{157}{See supra Section I.B.}
Including a disclaimer, however, precludes neither reputational harm to a candidate nor impacts on an election. Even if a deepfake’s falsehood is made apparent, evidence suggests that certain viewers will still internalize and amplify the message it conveys.\textsuperscript{158} Thus, a disclaimer alone is insufficient to render a deepfake harmless when other features of its publication and distribution increase its danger. Furthermore, subsequent bad actors can easily remove many forms of disclaimers, such as a simple text caption.\textsuperscript{159} Therefore, though the original post may declare that a deepfake contains falsehoods, subsequent posts—hundreds or even thousands of which may reach a dramatically larger audience than that of the original poster—\textsuperscript{160}—may not. For this reason, many statutes require stricter forms of disclaimers, such as digital watermarks, which are both more successful in preventing deception and more difficult to remove.\textsuperscript{161} Yet, even more sophisticated watermarks may be removed by subsequent sharers.\textsuperscript{162} Therefore, stricter disclaimer requirements do not solve all problems.\textsuperscript{163}

Finally, under some states’ laws, deepfake creators can essentially free themselves of liability, despite the likely negative effects of their content, simply by adding a disclaimer.\textsuperscript{164} Following that, subsequent sharers often have no responsibility to declare the inaccuracy of a deepfake they repost.\textsuperscript{165} Even the expansive House bill, which attaches liability to those who “alter” a disclaimer and therefore reaches culpable subsequent sharers, has its oversights. By not amending or removing the disclaimer, the subsequent sharer can satisfy their burden to avoid liability despite knowingly reposting a harmful deepfake.\textsuperscript{166} Therefore, a creator can post a deepfake with a disclaimer, a subsequent sharer can knowingly repost it, and neither would be subject to liability. Yet, even when a deepfake has a disclaimer, there is still real damage

\textsuperscript{158} See Harwell, supra note 29.

\textsuperscript{159} See Harris, supra note 156, at 117.

\textsuperscript{160} See infra note 166 and accompanying text.

\textsuperscript{161} See Langa, supra note 30, at 789.


\textsuperscript{163} Additionally, by priming viewers to believe that deepfakes will be accompanied by a standardized disclaimer, these requirements may actually have a negative impact: viewers will believe that any content without said disclaimer is legitimate, making deepfakes that are not accompanied by a disclaimer even more influential. Langa, supra note 30, at 789. Balancing this concern with the concern that fostering skepticism will lead people to distrust true media will likely prove difficult. But, hopefully, by assigning liability when appropriate and deterring bad actors, some of the worst effects of this dilemma will be mitigated.

\textsuperscript{164} See supra notes 57–58 and accompanying text.

\textsuperscript{165} See e.g., Act of July 23, 2023, ch. 360, 2023 Wash. Sess. Laws 1892 (to be codified at WASH. REV. CODE § 42).

\textsuperscript{166} Langa, supra note 30, at 789; Kocsis, supra note 145, at 643.
from its spread, since evidence shows that people will continue to believe a discredited deepfake if it affirms their preconceived notions.\(^{167}\)

2. Subsequent Sharers

Another policy consideration is whether to limit liability to the original publisher of a deepfake or extend liability to subsequent sharers. The House’s bill covers those who publish deepfakes without the necessary disclaimers and those who alter a deepfake to “remove or meaningfully obscure” the disclaimers.\(^{168}\) This is in stark contrast to some state statutes that have already been enacted, which typically assign liability only to the creator.\(^{169}\)

The case for extending liability beyond the creator is clear. Subsequent sharers can have a significantly larger deleterious impact than the original poster, spreading a deepfake to exponentially more people.\(^{170}\) Furthermore, when combined with other ill-advised elements, these statutes may unintentionally give rise to situations that attach no liability to anyone at all. Some statutes waive liability if a disclaimer accompanies a deepfake. If liability is limited to initial posters, subsequent sharers are left free of legal consequences for posting a deepfake without one. As a result, if the original creator attaches a disclaimer but subsequent sharers do not include it, the victim of the deepfake would have no recourse even though the damage caused could still be expansive.

Conversely, one benefit of limiting liability is avoiding the complication of trying to rein in culpable subsequent sharers without chilling innocent sharers’ speech. Another is avoiding the exceedingly arduous task of pursuing all sharers, an additional burden on top of the existing difficulty of tracking deepfake publishers.\(^{171}\) Though this Note’s approach addresses the innocent-sharer problem, the pursuit of subsequent sharers is likely to remain an issue given the ever-improving strategies to avoid detection.\(^{172}\) Additionally, an individual bringing claims against subsequent sharers may also face an uphill battle to establish the causation and redressability sufficient to establish standing.\(^{173}\)

\(^{167}\) See Harwell, supra note 29.


\(^{169}\) See e.g., TEX. ELEC. CODE § 255.004(d) (West 2023).


\(^{171}\) Langa, supra note 30, at 793.


\(^{173}\) Cf. Zipkin, supra note 148, at 194 n.67 (discussing standing in the context of gerrymandering and Elections Clause claims).
3. Reasonable Viewer

Last, baked into many of the statutes’ definitions of deepfake liability is a reasonable viewer standard: would the modified content in question deceive a “reasonable” person?\(^\text{174}\) If not, then the publisher is not liable. But there is good reason to think that such a standard is insufficient, especially in the context of politically oriented deepfakes.\(^\text{175}\)

First and foremost, this normative approach ignores the fact that any change in behavior at the ballot box is a harm worth curtailing.\(^\text{176}\) Why should it matter whether a viewer is not considered a reasonable person if they are deceived into changing their vote or convincing others to do so? After all, their vote counts just the same.

Furthermore, creators of politically oriented deepfakes may intentionally aim their work at unreasonable viewers. Studies have shown that an online community based around a particular political ideology tends to insulate its members from information that contradicts that ideology and feed members more information that confirms their beliefs.\(^\text{177}\) In other words, “intentionally false information . . . is often accepted and circulated.”\(^\text{178}\) Therefore, what might not deceive a reasonable audience may readily permeate such a partisan (and, perhaps, unreasonable) community. For example, scholars have documented the potential for extremist groups to harness deepfakes to gain support for violent movements or perpetuate anti-immigrant sentiments.\(^\text{179}\)

Current deepfake laws are far from ideal. These laws go both too far—to the point of chilling protected speech—and not far enough, failing to address the effects of some of the most harmful deepfakes. However, this Note does not advocate for a return to a landscape without any deepfake regulations. Rather, this discussion is meant to highlight the weaknesses of current laws and encourage reforms that strengthen enforcement capabilities and secure solid

---


175. It is also unavoidable that what is considered reasonable for a viewer will evolve over time with changing technology and increased awareness of that technology. Ice, supra note 103, at 437.

176. See Langa, supra note 30, at 781.


178. Schroeder, supra note 177, at 1099–1100.

constitutional footing. Part III proposes one such reform: a foreseeable harm standard.

III. THE BENEFITS OF A FORESEEABLE HARM APPROACH

In order to address many of the looming constitutional and policy concerns surrounding deepfake laws, future legislation should incorporate a foreseeable harm standard of liability, measured via a totality-of-the-circumstances test.¹⁸⁰ This approach allows for the consideration of several factors that may contribute to a deepfake’s harmful effects, rather than relying on discrete, poorly fitting rules. Among the factors that triers of fact could consider are: (1) the quality of the deepfake; (2) where the deepfake is distributed, such as via text message versus a social media post; (3) who receives the deepfake, keeping in mind specifically susceptible populations; (4) who posts the deepfake, because a newscaster or politician may have more credibility than, for example, a Redditor; (5) when the deepfake is published; (6) the presence of a disclaimer; (7) the type of disclaimer; and (8) whom the deepfake is meant to depict—directly or indirectly—and what it depicts about that person. Ultimately, the goal of a foreseeable harm standard is to determine whether the person that posted the deepfake should have foreseen the consequences of their actions. This approach would provide more protection for potential victims of deepfakes while also avoiding infringements of creators’ and sharers’ First Amendment rights.¹⁸¹

A. Benefits of a Foreseeable Harm Standard

1. Reasonable Viewer

Perhaps most importantly, a foreseeable harm standard would not have the same blind spots as the reasonable viewer standard. As discussed, not only is there a possibility that even crude deepfakes will deceive “unreasonable” audiences,¹⁸² politically oriented deepfakes are oftentimes intended to deceive these very audiences.¹⁸³ Therefore, attaching liability only when a deepfake

¹⁸⁰ Several formulations of this standard exist. Cf. infra Section III.B.2. Future legislation should, as a starting point, adopt some form of the following: “A defendant will be liable for any harm that, at the time of the defendant’s acts which constitute publication, would have been a foreseeable result to a reasonable person in the defendant’s position.”

¹⁸¹ Cf. Langa, supra note 30, at 781.

¹⁸² See Harwell, supra note 29 (“[W]hen [the truth of the alteration] comes to light, people just don’t care . . . . They say ‘it could have been true’ or ‘nonetheless, it reflects who the person really is.’” (alteration in original) (quoting Stanford University researcher Becca Lewis)).

¹⁸³ Supra Part II; see also Edward Lee, Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance, 70 AM. U. L. REV. 913, 1015–17 (2021) (demonstrating how politically oriented subreddits allow discriminatory content moderation, likely leading to bias).
would deceive a reasonable viewer ignores the fact that harm still follows from the dissemination of a deepfake that may not pass this standard.184

Unlike the reasonable viewer standard, the foreseeable harm standard would capture a creator or publisher of a deepfake who intentionally targets susceptible viewers and thus could predict the effect that their content would have.185 Furthermore, as scholars have noted, these publishers may not have any expectation of convincing someone who does not already believe the message of the deepfake; rather, the goal may be to simply “reinforce existing beliefs and get people more entrenched in those beliefs.”186 Thus, the publisher need not deceive a reasonable viewer to achieve their intended outcome: “the falseness barely matters . . . . [P]eople who believe the message already aren’t looking for a counternarrative: They just want confirmation that they were right all along.”187 Studies showing that “familiarity” increases believability further support this point.188 By seeking out politically persuadable and ideologically insular subpopulations, the deepfake publisher is more likely to find a susceptible audience. An objective of reinforcing messaging to certain populations would still make this harm, and any secondary harm resulting from it,189 worth remedying. Thus, a foreseeable harm standard would attach liability to these publishers.

2. Context-Blind Restrictions

A foreseeable harm standard would also avoid the danger of prohibiting a certain type of speech by any person, at any time, in any place. Rather than the categorical, context-blind rules that the Court warned against in Alvarez,190 a foreseeable harm standard is flexible enough to consider each of these factors, among others.

For example, *where* a deepfake is posted would affect how foreseeable the downstream harm would be to a candidate’s reputation or to a voter’s behav-

---

184. Cf. Anne Pechenik Gieseke, Note, “The New Weapon of Choice”: Law’s Current Inability to Properly Address Deepfake Pornography, 73 VAND. L. REV. 1479, 1484–86 (2020) (showing that, by putting deepfakes in a group like the “r/deepfake” group, they will be spread elsewhere and may be more likely to deceive viewers in that new context).

185. For example, after posting a deepfake that depicted President Obama shaking hands with a foreign leader, which was reposted by several thousand others, Representative Paul Gosar’s response was that it was the fault of the “dim witted” people that believed it, and that “[n]o one said this wasn’t photoshopped.” Harwell, supra note 29.

186. *Id.* (quoting Darren Linvill, a professor at Clemson University).

187. *Id.*

188. Vaccari & Chadwick, supra note 38, at 2 (arguing that viewers are more likely to accept misinformation if it conforms with familiar statements).

189. See, e.g., Pechenik Gieseke, *supra* note 184, at 1485–86 (describing how deepfakes spread from Reddit to other places).

ior. A creator texting a deepfake to a friend, arguably equivalent to “whisper[ing] within a home,” would foresee less harm than they would if they posted the same deepfake to an online extremist political group with thousands of members.

Likewise, who posts a deepfake would clearly enhance the foreseeable harm, given the poster’s credibility and the size of their audience. We have already seen how highly visible people sharing deepfakes—even if poorly made—can increase their impact, such as the deepfakes posted by Congressman Gosar, former President Trump, or Governor DeSantis.

Foreseeable harm would also take into account when someone posts a deepfake. Here, too, a totality test avoids adopting a hardline rule, like those employed in the BCRA and some existing deepfake laws. While it is true that the closer to an election a creator posts a deepfake, the more weight it may carry in the balancing of circumstances, the timing would not be dispositive, avoiding arbitrary distinctions. Instead, courts would weigh the timing of a deepfake’s distribution in the context of other factors. For example, a deepfake posted sixty-one days before an election, but otherwise particularly deceptive and harmful, may still give rise to liability.

In fact, considered together, Justice Kennedy’s concern for context-blind restrictions and Justice Breyer’s recognition of the potential for false speech related to political campaigns (a category that inherently encompasses political deepfakes) to impact voter behavior illustrate why the current patchwork approaches are an attempt to fill a circular hole with a square peg. The current laws are primarily aimed at determining how convincing the deepfake is. However, the more important question to ask is whether the person posting the deepfake acted in a way which they could reasonably have foreseen was

192. See Dhruva Krishna, Deepfakes, Online Platforms, and a Novel Proposal for Transparency, Collaboration and Education, RICH. J.L. & TECH., Spring 2021, at 1, 34. And the “where” can extend to specific communities within a single platform.
193. Barrett, Jr., supra note 38, at 635.
194. Harwell, supra note 29.
197. Supra notes 55–59, 139–142 and accompanying text (describing statutes that regulate speech within a specified number of days before a general election).
198. And even this proposition is likely not as black-and-white as one would think, at least in the context of changing voter behavior. A deepfake posted hours or even days before an election may in some instances have less damage than those posted further in advance if its false message didn’t have time to reach a mass audience.
199. Supra note 118 and accompanying text.
likely to harm a candidate’s reputation or impact an election. Take, for example, MIT’s deepfake of President Nixon. Using the best available technology, the MIT researchers created a deepfake that could have deceived a reasonable viewer more readily than most any other deepfake. Yet, MIT’s efforts were part of a project to educate the community on the extent of new AI capabilities. The way it distributed this deepfake, therefore, was less likely to have a harmful impact, and the creators knew as much.

That does not rule out the possibility of other people sharing MIT’s video in a different context and potentially deceiving audiences. Though this could create liability for those sharers, a court would probably not retroactively hold MIT liable, too, under the standard this Note proposes. To be clear, a deepfake can still give rise to liability even if it is unlikely to deceive by itself. Each assessment is hyper-contextual, asking when the user posted the deepfake and whether its nefarious purposes and effects were foreseeable.

3. Mens Rea

Deepfake laws should not impose liability on the “careless” but innocent speaker. Again, the intent of an individual who creates a deepfake is not debatable; certainly, after synthetically modifying an image or video, the creator did not genuinely believe that the content they shared was accurate. If the statute extends liability to all subsequent sharers who post a deepfake that would deceive a reasonable viewer, however, there is no guarantee that every individual will have had the same culpability. A foreseeable harm standard would allow for consideration of the specific circumstances in which each user shared the deepfake. For example, a careless distributor who does not realize they are sharing a deepfake does not foresee that their post could cause harm. Meanwhile, someone who purposefully removes a disclaimer before sharing a deepfake is more likely to foresee a deceptive and harmful effect.

200. Stockler, supra note 12. Thus, this standard is both overinclusive in some instances and underinclusive in others.

201. This is also relevant to satirical programs.

202. See supra Section II.A.iii.

203. Ice, supra note 103, at 434.

204. Supra Part II.

205. Thus, functionally, the “actual malice” requirement chosen by some legislatures will still be at play. The intentional or reckless behavior of any given sharer may be considered, along with the totality of their conduct, in order to determine liability. See Matthew Bodi, The First Amendment Implications of Regulating Political Deepfakes, 47 RUTGERS COMPUT. & TECH. L.J. 143, 162 (2021).

206. Deepfake laws are arguably a more realistic path to recourse for politicians than defamation, given the high bar to win such a claim. See Chesney & Citron, supra note 31, at 1793–94. The extent to which this remains true under a foreseeable harm approach, though, is unclear.
4. Subsequent Sharers

Another related consideration is whether to hold subsequent sharers liable or limit liability to the original publisher of a deepfake.\textsuperscript{207} A totality-of-the-circumstances test for foreseeable harm would not have to make this categorical call. Rather, it would take into account the knowledge of the subsequent sharer when evaluating the foreseeability of downstream harm. For example, if there is limited evidence that a subsequent sharer was aware the content was fake, the harm caused by their post would, naturally, be less foreseeable. This standard allows for flexibility in pursuing legal action against subsequent sharers who are particularly culpable—such as those who intentionally remove watermarks, publish deepfakes to susceptible populations, or take other steps to increase the likelihood of harm—and avoids situations where victims have no recourse despite significant harm. But the standard does so without chilling innocent speech by the careless sharer.

Of course, there would still be problems of proof in demonstrating what a sharer knew or what actions they took—for example, did they see a disclaimer and remove it?\textsuperscript{208} However, rather than viewing this as a bug, it can be seen as a feature: a check on chilling speech. As explained above, if there is no proof of the sharer’s culpability, a court will not hold that sharer liable.\textsuperscript{209} But with sufficiently serious evidence of culpability in the totality of the circumstances, subsequent sharers would not be able to wipe their hands clean of wrongdoing.

5. Disclaimers

A foreseeable harm standard would also allow for a more flexible approach to disclaimers by considering both whether a disclaimer was attached and the type of disclaimer attached.\textsuperscript{210} For example, a text caption can be easily

\textsuperscript{207} Compare, e.g., TEX. ELEC. CODE ANN. § 255.004 (West 2023) (only reaching the creator of a deepfake), with DEEPFAKES Accountability Act, H.R. 5586, 118th Congress (2023) (attaching liability to both deepfake creators and those subsequent sharers who “remove or meaningfully obscure” disclaimers).

\textsuperscript{208} Ice, supra note 103, at 433.

\textsuperscript{209} See Section III.A.3.

\textsuperscript{210} Another possible criticism of a system of liability based on the mere presence or absence of a disclaimer is that requiring a disclaimer is a form of compelled speech. Cf. Quentin J. Ullrich, Note, Is This Video Real? The Principal Mischief of Deepfakes and How the Lanham Act Can Address It, 55 COLUM. J. L. & SOC. PROBS. 1, 30 (2021) (arguing that this criticism is misplaced if the requirement leaves exceptions for types of speech that are less likely to be harmful). The flexibility of a foreseeable harm standard allows for a publisher’s inclusion of a disclaimer to reduce the chance that they are found liable without mandating a disclaimer in order to avoid liability.
removed by subsequent sharers down the line.\footnote{Pesetski, supra note 141, at 528; Coldewey, supra note 162.} By contrast, a digital watermark is harder to simply crop out of an image.\footnote{See Theo Golden, A British TV Network Is Facing Criticism for Airing a Deepfake Version of the Queen’s Christmas Speech, Where She Mocks Harry and Meghan for Moving to Canada, BUS. INSIDER INDIA (Dec. 26, 2020, 12:25 AM), https://www.businessinsider.in/tech/news/a-british-tv-network-is-facing-criticism-for-airing-a-deepfake-version-of-the-queens-christmas-speech-where-she-mocks-harry-and-meghan-for-moving-to-canada/articleshow/79956689.cms [perma.cc/S4C2-J877].} Thus, it would likely limit deception even after subsequent sharing and is more likely to cause a viewer to heed the manipulated nature of the content. The type of disclaimer a creator chooses to attach, and whether the creator chooses to attach a disclaimer at all, would change the assessment of foreseeability.\footnote{The ability to consider what type of disclaimer a deepfake publisher attached, if they attached one at all, is another benefit to pursuing a claim under a deepfake law as opposed to a traditional defamation claim. Any declaration that the content is false would render a defamation claim moot. See Kugler & Pace, supra note 93, at 630. Meanwhile, liability may still attach under a defamation law if the disclaimer is not sufficient to dispel harm down the line.} Rather than a binary distinction, between the presence or absence of a disclaimer or between a textual disclaimer and a digital watermark, to determine liability, a foreseeable harm standard would consider the existence or nature of the disclaimer as only one aspect in the totality of the circumstances.

Even identical disclaimers will not always have an equivalent impact. In MIT’s Apollo project, a digital watermark or auditory disclaimer was not necessary to convey the video’s authenticity (or lack thereof) due to the surrounding context of its creation.\footnote{See supra notes 12–13 and accompanying text.} Meanwhile, a deepfake that is accompanied by a watermark disclaimer, but that is shared to a subgroup known to possess the ability to remove such watermarks\footnote{See Coldewey, supra note 162; see also supra notes 208–212 and accompanying text.} and spread the manipulated content to other communities, would have greater foreseeable harm. This would be true even if the creator included what is normally a highly effective disclaimer.

6. Who and What Are Depicted?

A foreseeability test would also consider what the deepfake depicts and whom it targets. The notoriety of the candidate, the stage of the campaign, and the relative importance of the election are all factors that could increase the audience of deepfake targets. Certain candidates also inspire more targeted partisan opposition, which in conjunction with extremism and the familiarity effect may increase the likelihood of deleterious outcomes.\footnote{For example, Representative Pelosi has long been the center of Republican hatred and misinformation campaigns. See Annie Karni, Catie Edmondson & Carl Hulse, Pelosi, Vilified by Republicans for Years, Is a Top Target of Threats, N.Y. TIMES (Nov. 10, 2022), https://www.nytimes.com/2022/10/30/us/politics/pelosi-attack-republican-threats.html[perma.cc/VM5P-R5YA]; Annie Karni, Malika Khurana & Stuart A. Thompson, How Republicans Fed a Misinformation...
Additionally, this approach could account for whether the deepfake depicts the candidates themselves, someone associated with a campaign, or something else entirely but that is still intended to impact viewers’ impression of a campaign. Presumably, if a creator makes a deepfake to directly target a specific candidate, they should foresee its harmful impact. There is certainly still a possibility, however, that a deepfake that manipulates something other than a candidate or their campaign directly could still significantly harm that candidate. Of the current laws in place, some appear to restrict liability only to sharers of deepfakes that depict candidates themselves, while others do not require that a candidate be depicted so long as the deepfake is meant to alter a viewer’s impression of the campaign or election. Thus, a deepfake placing a candidate’s spouse or campaign manager in a compromising context would escape the grasp of some current laws. At the same time, all of the current laws require that the deepfake depict a person. A deepfake falsely portraying an object or event would fall outside their ambit. Yet, the harm that misinformation like this could cause—including, in the most extreme case, altering an election—should be foreseeable.

B. Support for a Foreseeable Harm Approach

Using United States v. Alvarez as a barometer, this Section explains the constitutional support for the foreseeable harm standard. In addition to sidestepping some of the pitfalls of the Stolen Valor Act, this approach also finds support in Justice Breyer’s concurrence. Finally, this Section draws analogies

---

217. See Ice, supra note 108, at 439 (discussing how there are more damages than just reputational ones to the person depicted). It is beyond the scope of this Note to wholly address the necessity of broadening the scope of deepfakes for which victims may seek recourse, but it is worth addressing how such an expansion would fit into the framework of a foreseeable harm standard.

218. See supra Section I.B.


220. See supra Section I.B.

221. Deepfakes of objects or events directly targeted at a campaign could hypothetically depict, to give a few examples, controversial documents, campaign officials destroying property, or a campaign vehicle engaging in illegal conduct. More indirect impacts could stem from deepfakes that negatively impact an incumbent administration, such as a series of recent deepfakes that ostensibly depicted explosions near the Pentagon and the White House. See Shannon Bond, Fake Viral Images of an Explosion at the Pentagon Were Probably Created by AI, NPR (May 22, 2023, 6:19 PM), https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai [perma.cc/ET39-57TW]. In addition to undercuts the reputation of our national security apparatus, these images also temporarily caused stock market dips. Id.
between different “reasonable person” standards and identifies the proper positioning of deepfakes in this framework.

1. The Alvarez Concurrence

In his concurrence, Justice Breyer discussed some types of justified speech restrictions, as well as the limitations that allowed them to satisfy the requisite level of scrutiny. One example is false claims of terrorist attacks. In order for liability to attach in that context, there must be “proof that substantial public harm [was] directly foreseeable, or, if not, [the claim must] involve false statements that [we’re] very likely to bring about that harm.” Justice Breyer also briefly touched on claims of trademark infringement; these claims require a finding that the infringement is “likely to dilute the value of a mark” or cause confusion. Building upon the requirement that speech restrictions be limited to contexts in which the greatest degree of harm may arise, this use of a foreseeable harm liability standard suggests that it also matters how foreseeable the specific harm in question was to the perpetrator.

Trademark infringement claims, along with impersonations of a public official, which Justice Breyer also considered, are surprisingly analogous to deepfakes. In both contexts, the tortfeasor tries to capitalize on the inherent authority and believability of the brand they are attempting to resemble, or the role they are pretending to hold, to alter the target audience’s perception of them. As Justice Breyer noted, this is more than “mere speech.” Impersonating a public official, like publishing a deepfake, is an act meant to gain credibility—more credibility than is attainable by simply making a false statement. And this attempt to misappropriate credibility should have consequences. If a creator invests in making a deepfake believable, they should be able to predict that a viewer will actually act on that deepfake and change their behavior. Thus, the more likely a speaker’s conduct is to change their audience’s behavior in accordance with the false statement, the more appropriate liability is for the perpetrator.

---

222. See supra note 103 for a discussion of Justice Breyer’s and Justice Kennedy’s differing views on the appropriate tier of scrutiny applied in these cases.


224. Id. at 735–36.

225. This is a different similarity than that which Justice Breyer considered between trademark infringement and the Stolen Valor Act.

226. See Sunstein, supra note 91, at 422; Ice, supra note 103, at 434 (describing the authority and believability garnered by deepfakes); Alvarez, 567 U.S. at 735 (Breyer, J., concurring) (explaining the role of seeking credibility when impersonating a public official).

227. Alvarez, 567 U.S. at 735 (Breyer, J., concurring). As for how Justice Breyer distinguished between conduct that can be prohibited and “mere speech,” in the context of impersonating a public official, he stated that these statutes “typically focus on acts of impersonation.” Id. The act of creating a deepfake or altering its disclaimer would fall comfortably into the same category.
2. Other Applications of the “Reasonable Person” Standard

Beyond the First Amendment examples listed in Justice Breyer’s Alvarez concurrence, it is important to recognize that foreseeable harm is far from a novel legal standard. Foreseeability has a well-established history in tort and contract law. Notably, though the reasonable person standard advocated for here is not the same as the reasonable person standard incorporated into current deepfake laws, it is still a form of the reasonable person standard. In fact, it simply flips the current standard on its head. Rather than asking if a reasonable person—the viewer—would be deceived by the deepfake, my proposal asks whether a reasonable person—this time, the publisher or sharer of the deepfake—would have known that harm would result. This version of the reasonable person standard is meant to ascertain whether “an ordinary person in [the] same circumstance would have reasonably acted in the same way.”

It is an evaluation of the tortfeasor’s behavior, rather than a measure of the victim’s susceptibility.

Current laws, which ask if the deepfake would deceive a reasonable viewer, are aligned with the common features of harassment claims, which ask if the defendant’s conduct would have made a reasonable person feel harassed. Harm arising from a deepfake, however, is distinguishable from harassment claims and should not be governed by the same liability standard. In harassment cases, in addition to determining whether the claimant actually found

---


229. For a discussion of a similar approach to assigning liability in defamation cases, see Alex B. Long, All I Really Need to Know About Defamation Law in the Twenty-First Century I Learned from Watching Hulk Hogan, 57 WAKE FOREST L. REV. 413, 459 (2022) (discussing how the Texas Supreme Court in the case of New Times, Inc. v. Isaacks, 146 S.W.3d 144, 162 (Tex. 2004), asked if “the publisher either kn[e]w or ha[d] reckless disregard for whether the article could reasonably be interpreted as stating actual facts”). This Note argues that, in the context of politically oriented deepfakes, the Texas Supreme Court’s analysis should take the next step to exclude the requirement of a reasonable viewer so long as the publisher could themselves reasonably foresee the harm to follow.


231. See, e.g., Harris v. Forklift Sys., Inc., 510 U.S. 17, 21 (1993) (reaffirming the standard that, to be actionable, workplace discrimination must be both subjectively and objectively “hostile or abusive” (citing Meritor Sav. Bank v. Vinson, 477 U.S. 57, 64–65, 67 (1986))).
the conduct offensive, courts ask whether a reasonable person would have found the conduct offensive.232

On the other hand, when determining a deepfake’s harm under a reasonable foreseeability standard, we would not ask whether the viewer—the deceived—was reasonably offended. Nor would it matter whether the deceived reasonably feared a future harm to themselves, which is a part of establishing an assault claim (another type of tort that utilizes the equivalent of a reasonable viewer, rather than reasonable perpetrator, standard).233 We still consider future harm in the deepfake context when assessing whether the deception of the viewer harmed a candidate’s reputation or unduly influenced an election. But this is not harm to the deceived. Rather, the deceived is used as an agent to harm a candidate or the election process. The deceived is not seeking redress for themselves. Instead, the person whose reputation was harmed, or who was impacted by an election interference scheme, is seeking redress for the harm done to them by the perpetrator—using the deceived as an agent. Thus, whether the deceived was a reasonable person should not impact liability; either way, the ultimate victim, the subject or target of the deepfake, was still harmed. So, much like Justice Breyer’s impersonation and infringement examples, courts should be asking if the perpetrator acted reasonably—if they should have foreseen the result of their actions regardless of the reasonableness of the deepfake viewers.

Another important distinction is that, in harassment and assault, we must consider whether a reasonable person would feel harassed or threatened to determine if there was an injury; otherwise, people could make a claim at any time and subject defendants to punishment. In the deepfake context—much like incitement to crime234—the injury is not dependent on the viewer’s feelings being reasonable. Whether the deceived viewer was reasonable or not, the victim was harmed. Therefore, just as we ask whether a defendant charged with incitement should have foreseen the consequences of their words,235 we should ask whether a deepfake creator or distributor foresaw the harm, instead of asking whether the deception was reasonable.

The proposed House bill does contemplate a form of the foreseeable harm requirement. In the case of a deepfake depicting a deceased person, one of the


234. Foreseeable harm is also a recurring test in criminal law. For example, those states that determine coconspirator liability under the Pinkerton v. United States standard ask whether the harm was “reasonably foreseeable as a necessary or natural consequence” of their actions. Pinkerton v. United States, 328 U.S. 640, 647–48 (1946). Likewise, when considering whether a subsequent event severs the causal link from a defendant’s actions to the harm suffered, we ask whether the intervening cause was reasonably foreseeable to the defendant at the time of their misconduct. See, e.g., People v. Rideout, 727 N.W.2d 630, 633 (Mich. Ct. App. 2006) (citing People v. Schaefer, 703 N.W.2d 774 (Mich. 2005)).

requirements for liability is that the deepfake is “substantially likely to either further a criminal act or result in improper interference in an official proceeding, public policy debate, or election.”\textsuperscript{236} However, this provision is limited to one narrow category of deepfakes, and the proposed bill also retains several of the aforementioned troublesome elements, including the reasonable viewer standard.\textsuperscript{237} Even within this narrow category of deepfakes subject to the foreseeable harm standard, the bill speaks to foreseeable harm in terms of election interference, but not in terms of reputational harm to a candidate or campaign.\textsuperscript{238} Though the House’s proposed language differs from this Note’s solution, it demonstrates that Congress is aware of and open to a foreseeable harm test for deepfake liability down the road.

**CONCLUSION**

As deepfake technology continues to develop and cases percolate in the courts, there must be an effective, workable, and constitutional approach to assigning liability for the detrimental effects of politically oriented deepfakes. In place of the current patchwork system of often ill-fitting and potentially unconstitutional rules, a foreseeable harm, totality-of-the-circumstances test will allow for flexibility without unnecessarily curtailing protected First Amendment speech. Using foreseeable harm as the standard is both more narrowly tailored than current statutes and grounded in existing Supreme Court precedent. The House of Representatives’ proposed bill to address deepfakes takes the first steps toward a foreseeable harm standard, with, hopefully, a more full-scale adoption of this approach still yet to come.

\textsuperscript{236} DEEPFAKES Accountability Act, H.R. 5586, 118th Cong. § 1041(n)(1) (2023).
\textsuperscript{237} Id.
\textsuperscript{238} Id.